

Ending the Blame Game on Educational Inequity: A Study of “High Flying” Schools and NCLB

by

**Douglas N. Harris
Assistant Professor
Florida State University**

Education Policy Research Unit (EPRU)
Education Policy Studies Laboratory
College of Education
Division of Educational Leadership and Policy Studies
Box 872411
Arizona State University
Tempe, AZ 85287-2411

March 2006

EPSL | **EDUCATION POLICY STUDIES LABORATORY**
Education Policy Research Unit

EPSL-0603-120-EPRU
<http://edpolicylab.org>

Education Policy Studies Laboratory
Division of Educational Leadership and Policy Studies
College of Education, Arizona State University
P.O. Box 872411, Tempe, AZ 85287-2411
Telephone: (480) 965-1886
Fax: (480) 965-0303
E-mail: epsl@asu.edu
<http://edpolicylab.org>

This research was made possible by a grant from the Great Lakes Center for Education Research and Practice.

Ending the Blame Game on Educational Inequity:

A Study of “High Flying” Schools and NCLB

Douglas N. Harris
Florida State University

Executive Summary

One of the central purposes of public education is to provide opportunities for all children to learn and excel. Unfortunately, while gaps in educational outcomes have indeed improved substantially over the past half-century, poor and minority students are still well behind their more advantaged counterparts. There is also evidence that the positive trend has reversed course—that educational outcomes are now becoming even more inequitable.

Recent policy studies by the Education Trust and Heritage Foundation have tried to identify “high-flying” schools—schools that help students reach very high levels of achievement, despite significant disadvantages. This policy brief demonstrates three major problems with the findings of these reports. (1) Due to questionable methodological assumptions, the number high-flying schools is significantly smaller than the number reported in those studies; (2) The numbers in these reports are being misused in a way that that understates the significance of, and need to address, socioeconomic disadvantages; and (3) these reports fail to directly address the vast amount of evidence that inequity in educational outcomes is primarily due to students’ social and economic disadvantages.

It is therefore recommended that:

1. Policy makers continue the recent focus on measurable student outcomes, such as test scores, but redesign policies to hold educators accountable only for those factors within their control;
2. Policy makers take a comprehensive approach to school improvement that starts in schools but extends into homes and communities, and addresses basic disadvantages caused by poverty; and
3. All educational stakeholders acknowledge that educational inequity is caused by problems in both schools *and* communities—and avoid trying to blame the problem on schools alone.

Ending the Blame Game on Educational Inequity:

A Study of “High Flying” Schools and NCLB

Douglas N. Harris
Florida State University

Background

The achievement gap between students of various racial, social, and economic groups is large and growing. For example, between whites and African-Americans, the size of the achievement gap ranges from 29 to 37 percentile points. Between whites and Hispanics, the gap is 16 to 34 percentile points.¹ Strong signs suggest these gaps have worsened recently after decades of improvement.²

All parts of the political spectrum seem to agree that these educational inequities represent a significant problem. There is also strong evidence and agreement that students' social and economic disadvantages are substantial causes of the problem.³ Poor nutrition and illness cause students (a) to miss school more often and (b) to be less prepared to learn when they attend.⁴ Within the disadvantaged home, parents often have relationships with their children that are, emotionally and physically, less healthy.⁵ These unhealthy relationships are reinforced in part by economic pressures that induce conflicts between parents and children.⁶ The combination of these factors and other effects is shown to be worse as students remain in poverty for longer periods of time.⁷ Of course, many parents living in poverty are able to successfully navigate and avoid these potential problems, and some parents with high incomes are not great parents, but the general patterns described here are quite strong.

Perhaps the best evidence on students' disadvantages comes from a recent study of children when they first enter kindergarten. Because these students have not been in school, any observed inequity can only be attributed to family, community, and related factors that are outside of school control. This evidence suggests that the achievement levels of African-American kindergarteners are 34 percentile points below the levels of white kindergartners—roughly the same as students much later in their school careers.⁸ Again, the intention here is not to equate race with disadvantage, or disadvantage with poor parenting. The point is that alleviating the harmful effects of social and economic disadvantage is an important component of any effort to reduce educational inequity.

Of course, addressing disadvantages caused by family and community factors is not the only strategy for addressing educational inequity. Indeed, a common argument made in the policy arena is: Because the government has relatively little control over what goes on in the homes and communities of children, it has no choice but to focus efforts in the one place it has some control—public schools.⁹ One strategy is to try to make up for student disadvantages through extra resources. While the effects of such resources are positive for disadvantaged students on the average, some researchers have concluded that the effects are too small to be worth the costs.¹⁰ An alternative, and increasingly common approach, is for state and federal governments to use higher standards and accountability to induce school to do more with the resources they already have. On this point, evidence that some of these policies can improve educational equity exists, but other evidence suggests that they undermine good instruction. Therefore, as with the debate on resources and funding, the results are inconsistent.¹¹

What is clear, no matter how the evidence is interpreted, is that no single solution will solve the problem. Improving home and community environments would clearly help, but it is difficult (and not necessarily desirable) to try to control them. Conversely, schools are somewhat easier to control, but they may not be the primary source of the problem and they certainly are not the sole source of its solution. It seems evident that a comprehensive approach to educational inequity is necessary to substantially reduce it.

This conclusion would not seem to be very controversial, but, as the next section will show, some educational reformers appear to view the matter very differently. In particular, recent Education Trust and the Heritage Foundation reports suggested that the responsibility for educational inequity lies solely with schools. More significantly, the same view underpinning these recent reports—that schools are almost entirely to blame for educational inequity—is also a basic assumption now embedded in educational policy at both the state and federal levels.¹²

Adopted in 2001, the federal reauthorization of the Elementary and Secondary Education Act, commonly known as *No Child Left Behind* (NCLB), requires all students to achieve proficiency, as measured by standardized tests, in all subjects by the year 2014. In the meantime, schools must make Adequate Yearly Progress (AYP) towards that goal or face sanctions. To measure progress, schools must test students in all grades three through eight and the scores must be reported by racial and economic sub-groups. Moreover, all sub-groups must eventually become proficient. For equity purposes, this last point is potentially important: If all students were able to reach these proficiency objectives, then the gap will be not just reduced, but apparently eliminated.

There are many things to like about NCLB, especially its apparent ambition, its focus on measurable student outcomes, and its stated concern for the disparities in outcomes among different socio-economic groups. But the law suffers from the same flawed assumption as the Education Trust reports, implicitly placing all of the blame for educational inequity on schools. With NCLB, schools are judged based on the levels of student achievement rather than how much students learn in school. Therefore, even if a disadvantaged student enters kindergarten far below other students, and even if the school is very successful in helping the student learn, the school will still be punished if the student does not reach the proficiency cut off. This is not the only way that NCLB places responsibility solely on schools, but it is the most important.¹³

The “Recent Developments” section describes the Education Trust and Heritage reports and shows how they invite a false interpretation. Problems arise because the report’s limitations in the research methods and some related statistical issues, such as “regression to the mean” and use of test score “proficiency” definitions. These issues, discussed in the “Available Data” section, have important implications for both the Education Trust reports and the measures of proficiency in NCLB.

The “Available Data” section provides detail on the database used for the report’s analysis, the School-Level Achievement Database (SLAD) developed by the U.S. Department of Education—the same database used by Education Trust (ET) to generate its findings. A description of the database follows, explaining how it provided data for the alternative analyses, and offers an overview of its strength and weaknesses as a source of information on what is actually happening in schools. An analysis of these data

is provided in “Discussion and Analysis of Available Data” and, from this, the final section offers a series of recommendations for educators and policy makers.

Because this study is partly about the misinterpretation of other studies, it is important to be clear about the purposes and appropriate uses of material presented here. First, this is not a study of whether NCLB will be effective in reducing the achievement gap. While the data available in the SLAD are useful for the analyses presented below, they are not appropriate for identifying policy effects. In addition, this is not another broad-based attack on NCLB. As indicated earlier, the focus on measurable student outcomes—including the achievement gap—is an important positive step. At the same time, the law does make some fundamentally flawed assumptions, creating problems in its design that need to be addressed.

Recent Developments

High Flyers and No Excuses

The focus of the present study is on Education Trust’s 2001 report that identifies high-flying schools based on data regarding student achievement and student demographics.¹⁴ Specifically, the report defines “high-flying” schools as those that are both “high-performing” (above the 67th percentile in average state standardized test scores) and “high-poverty” (more than 50 percent of students are eligible for free or reduced price lunch). They find 3,592 schools that meet these criteria.

This number is problematic because it ignores the much larger number of schools that are unable to overcome student poverty, giving the impression that overcoming

poverty is relatively easy. The number 3,592 may seem large, but, as the next section shows, it is actually a small fraction of the high-poverty schools around the country.

A less obvious limitation is that the Education Trust definition does not require performance at a consistently high level—it requires high achievement in only one subject and considers only one grade and one year. As a result, it would call a school “high-flying” even if students could not read or do basic math. Moreover, it does not require that schools produce high achievement over time or in multiple grade levels. This leads to misidentification of high-flyers and overstatement of the total number, as shown in the analysis in the later sections.

In March, 2002, Education Trust followed this with additional analyses that used different definitions of high performance in an attempt to address some of these criticisms.¹⁵ They also try to minimize the problem with their earlier definitions, writing that “no single definition of high performance—or high-poverty or high-minority, for that matter—will work for all research purposes.”¹⁶ This is undoubtedly true, but it misses the point of the critique. Different definitions are appropriate under different situations, but some definitions of high performance should not be used except when absolutely necessary. To educators and education researchers, it is well known that individual test scores are unreliable measures of student achievement that vary dramatically from year-to-year and grade-to-grade even when school effectiveness is unchanged. Any definition that does not take this into account will likely yield misleading results no matter what type of research is being done.

The Education Trust report authors also write in support of their original performance definition that “we know from our own work in schools across the country

that the reforms that take hold in one subject and one grade level can provide the basis for improvements in other grades and subject areas.” This is almost certainly true, but schools that are improving should eventually achieve high scores in more than one subject, grade, and year. Without identifying schools that have improved in this way, it is difficult to learn how improvement takes place. In short, the performance definition in the original Education Trust report is ill-suited for the stated task.

Inviting Misinterpretation

It is easier to understand the origin of these methodological flaws when considering how these organizations view educational inequity and reform. Consider the words of Kati Haycock, Director of the Education Trust (ET). She asks, “How many effective schools do we have to see in this country before we conclude that it’s not about the kids?”¹⁷ One possible interpretation of this quote is that some students grow up under adverse circumstances, placing them at a disadvantage in their school activities. Therefore, it may not be “about the kids,” but rather about the conditions under which they live and grow. This interpretation is consistent with the research evidence.

But Haycock’s words invite an alternative interpretation. If we ignore the fact that harsh family and community conditions hurt children, then the choice is between blaming the schools and believing that some students are incapable of learning no matter what schools do. To see why, consider the foot-race analogy made by President Lyndon Johnson when he argued for affirmative action and compensatory education. Johnson said that undernourished students would lose the vast majority of the running races, not because the students or track coach failed to try hard enough, but because the students

were undernourished. Haycock's words imply that we should ignore the under-nourishment and other social and economic disadvantages.

The unfortunate result is that the Education Trust studies set up a false choice—a choice between blaming the students and blaming the schools. Given this choice, one can only blame the schools. And indeed, this is exactly what happened when the report was released:

“People who follow education issues have long known that some schools succeed with children from families with weak educational backgrounds. But it turns out [according to the recent Education Trust report] that it's not just a few, rare schools that succeed, it's thousands of schools . . . We'd better not hear that racist nonsense anymore.” Bill Evers, Research Fellow, Hoover Institution, *Brainstorm* NW Magazine, February, 2002.

According to Evers, you either believe that the schools are to blame or you believe in racist nonsense. But this view completely ignores the fact that family and community factors play a critical role. The belief that these factors are important is far from racism. Indeed, ignoring these family and community factors only reinforces the false view that some students are incapable. Unfortunately, the Evers quote is just one of many examples of how the Education Trust results have been interpreted.¹⁸

Heritage Foundation, No Excuses

The recent Education Trust reports share many similarities with a 1999 report published by the conservative Heritage Foundation, entitled, *No Excuses*.¹⁹ Its analysis started with approximately 400 schools brought to its attention from various sources, including state education agencies, think tanks, teachers' unions, and foundations. Like

the Education Trust report, the authors of the Heritage study narrowed this list to 125 schools that had high concentrations of poverty and high test scores. Its specific criteria were also similar—to be on the list, test scores had to be in the top-third of the state and at least 75 percent of the students had to be eligible for free and reduced lunch (instead of 50 percent in the Education Trust report). From this list of 125 schools, 21 were selected for site visits and further study.

The most significant problem with the Heritage report is that nearly all of the schools considered, while perhaps very effective, had unique resources or student populations that had little to do with the school's effort. For example, nine of the 21 schools had admission requirements that could exclude students who have received low test scores. Overall, a more careful analysis shows that only three of the 21 schools could be considered high-flyers.²⁰ Much could be learned from these schools, but the Heritage study masks the lessons of the analysis rather than learning from them.

In the foreword to *No Excuses*, Adam Myerson, then-Vice-President of Educational Affairs at the Heritage Foundation, states that some people would “dismiss such achievement as a fluke . . . the work of extraordinary heroes whose performance cannot possibly be held as a national standard” (p.2). Myerson is right that the high scores in these schools are no “fluke.” What he fails to recognize from his own information is that high performance of many schools can be explained substantially by systematic differences in family and school resources that are outside educators' control.

The NCLB Connection

The connection between the Education Trust and Heritage reports and No Child Left Behind (NCLB) is important to point out. In particular, these reports and the new

law all assume that schools are mainly, or even solely, responsible for educational inequity. In the case of the Education Trust reports, this appears to be a conclusion of the data analysis, but the discussion above shows the analysis only reinforces the authors' misguided assumptions. With NCLB, the same assumptions are revealed by the fact—not widely recognized—that schools are not actually punished or rewarded for what schools contribute to student learning. Instead, the law provides incentives for schools based on the percent of students who reach proficiency. This may sound reasonable; however, it completely ignores the vast differences in where students start—as documented by the research cited earlier on kindergartners. This means that many schools will be punished for family and community factors that are outside of their control—and therefore assumes that schools are solely responsible for inequity.

Methodological Issues

The false assumption that schools are primarily responsible for educational inequity is also reinforced by certain methodological limitations of the Education Trust reports and the federal law. These are related to two factors: regression to the mean and the use of proficiency definitions.

Regression to the Mean

Researchers assume that all measures are made up of two parts: (1) the true portion, or “signal,” which is the part of greatest interest, and (2) “noise.” Noise is assumed to be random in the sense that it is unrelated to the signal portion of the measure. In addition, the expected value of the noise for each individual is zero. This means that observed measure is different from the true value, but the direction and size of the difference are unclear.

One effect of statistical noise is called “regression to the mean.” For instance, suppose you flipped a coin ten times and obtained nine “heads” and one “tails.” Such a pattern cannot go on forever. If you continued flipping, the average number of heads would gradually converge to 50 percent. More generally, if we repeat any measure, the average will tend to shift towards the expected or mean value. It is therefore easy to see why the concept is called “regression to the mean.”

This effect also occurs with schools and test scores. If a school achieves a very high score, it is likely that some, though certainly not all, of this high performance is caused by positive noise—factors outside of the school’s control but that nonetheless affect measured student test scores. Because noise is considered random, it is unlikely that the same school will experience positive noise for all other tests. Other attempts will likely produce lower scores unless the school is truly exceptional.

Unfortunately, some recent studies show that the signal-to-noise ratio of standardized test scores is very low, implying that the role of regression to the mean can be quite large.²¹ As a practical matter, this means that adding additional test scores (e.g., test from additional grades, subjects, and years) could significantly change measured levels of achievement in many schools. Because such additions reduce the effect of regression to the mean, and help us come closer to the real achievement levels, it is important that the additional data be included.

The effect of statistical noise is further complicated when schools are separated into low-poverty and high-poverty categories—as is the case in the Education Trust study—because the two groups have a different expected score. A concrete example may help to illustrate. Consider a typical high-poverty school, School H, and a typical low-

poverty school, School L. If there were no noise, School H would achieve the 40th percentile and School L would reach the 70th percentile. While the expected effect of noise is zero, suppose that each school has a 20 percent chance of receiving positive noise equal to 30 percentile points (i.e., noise that raises reported scores above true scores) and a 20 percent chance of experiencing equally-sized but negative noise. Now, suppose that in year one, School H experiences positive noise and therefore reaches the higher-than-expected 70th percentile, and School L experiences no noise, and therefore reaches the expected 70th percentile. Both schools are high-performing according to the definitions used in the Education Trust analysis.

However, the odds of this happening again are slim. There is only a 20 percent chance that School H will experience positive noise again, so the school will probably switch from the high-performing group to the low-performing group. School L, in contrast, has an 80 percent chance of remaining high-performing because there is only a 20 percent chance that it will experience negative noise large enough to decrease its percentile below the cut score.

What this means for the analysis of achievement gaps is: (1) all schools that appear high-performing at any given point in time may actually be average or below; and (2) just as importantly, this *false identification is much more likely to occur with high-poverty schools*. The results in the “Evaluation of Available Data” section below confirm this effect and also demonstrate why it is essential to use a substantial number of scores when trying to identify school performance.²²

Proficiency and “Cut Scores”

There are many different types of standardized tests and many ways to report them. One general approach reports school test scores as averages of the scores from individual students. Such measures incorporate the performance of all students, and therefore improvement by any given student, no matter their initial level of achievement, appears as a slightly higher school average.

An alternative approach is to create a “cut score” and use it to distinguish between “proficient” students who score above the cut and “non-proficient” students who score below the cut. The purpose of this approach is to establish a minimum benchmark that all students are expected to attain. This is certainly a reasonable means to understand the overall level of achievement among broad groups of students. These cut scores, however, are problematic when used for the sake of school accountability. One problem is that accountability systems using cut scores create an environment where schools focus all of their attention on the students who are just below or just above the cut score because the other students are likely to remain in the same category even if the school devotes little attention to them. A second problem, as indicated earlier, is that even a highly effective school might not be able to help a student who starts off far behind to achieve at the same level as other students.

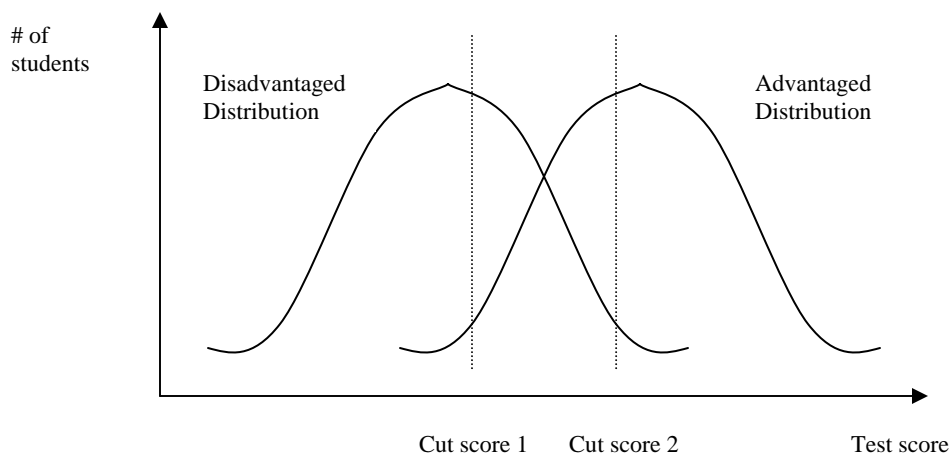
One prominent education scholar, Richard Rothstein, writes that the specific cut score chosen for analysis purposes causes “great mischief” with the measure of achievement.²³ He argues, for example, that an extremely low cut score is likely to be reached by high percentages of students in all groups, making the achievement gap seem small. Conversely, very low percentages of students in all groups will reach extremely

high cut scores, resulting in a similarly low achievement gap. As a result, Rothstein writes, “critics can make the test score gap seem extraordinarily large if they define proficiency about halfway between the average score for blacks and the average score for whites.”²⁴

This is illustrated in Figure 1 below which displays realistic test score distributions for disadvantaged and advantaged students. The bell-shaped distribution to the left has a lower test score mean and reflects the distribution of disadvantaged students. The other similarly shaped curve has a higher mean score and reflects advantaged students. Two cut scores are also shown. At the first, nearly half of the disadvantaged students are proficient, but at the second, almost none of them are.

The two score distributions and two cut scores illustrate why the two groups of students are affected differently by changes in the cut score. Specifically, a small change in cut score 1 will have a larger effect on the proportion of disadvantaged students passing the exam. For cut score 2, the opposite is true; now, the advantaged group is affected more. More generally, when a policymaker moves the cut score closer to the intersection of the two distributions, the gap will appear larger. While this requires other assumptions, it does illustrate and clarify Rothstein’s point that the cut score causes “great mischief.”²⁵

Figure 1: Effect of the Cut Score on Different Student Groups



As we will see below, this has important implications for NCLB because the law encourages states to define proficiency at a very low level (closer to cut score 1), making the achievement gap seem small.

Available Data

The recent Education Trust reports are based on a database created by the U.S. Department of Education, the School-Level Achievement Database (SLAD). When used properly, the SLAD is potentially useful for identifying high-flying schools. It is important, however, to understand both the promise and the limitations of this tool. The following sections offer a general description of the database and provide an overview of SLAD's general advantages and limitations.

The SLAD has a total useable sample of 62,074 schools (74 percent of all public schools in the country) that enroll 36 million students (78 percent of the total). For most schools in the SLAD, information is included about the percentage of students in various

racial and ethnic categories and the percentage of students eligible for free or reduced lunch.²⁶ Information about school location (urban, rural, etc.), school type (charter, magnet, and traditional public), and school grade levels (elementary, middle, and high schools) is also included.

The data for this new analysis study are from the years 1997-2000. More recent data are available; however, there are two reasons for using the older data. First, by using data before NCLB was adopted, it is possible to avoid the issue of whether NCLB has affected test scores, positively or negatively. Second, the older data include those used by Education Trust in their 2001 report, making it easier to compare results across the two studies.²⁷

For these years, test scores are reported in the SLAD for all U.S. states, except Iowa, South Dakota, and West Virginia.²⁸ In some of the 47 included states, there is no standardized test at the high school level; therefore, such schools are also excluded. Every state gives a different achievement test and reports these results in different ways.²⁹ One useful feature of the SLAD is that it includes data from multiple cut scores in many states, making it possible to illustrate some of the points made in the previous section.

Schools are divided into three levels: elementary, middle, and high school. Where possible, the following grades were chosen: grade five for elementary school, grade eight for middle school, and grade 12 for high school. In cases where scores were not available at these grades, the next lowest grade was used (e.g., grade four was used instead of grade five). The analysis compares schools within the same general grade levels, e.g., elementary schools are only compared with other elementary schools, not middle schools.³⁰

For some states, information is available for multiple grades, years, and subjects. In some states, at least eight test scores are available for each school—two subjects and two grades for two consecutive years. This “multiple scores” sample includes 18,365 of the 62,074 schools. Of these, 14,124 are elementary schools, 4,241 are middle schools, and none are high schools.

Strengths and Limitations of the SLAD

The SLAD is apparently the only database that comes close to providing school demographic and achievement information for all schools in the U.S. Other data sets provide richer information for small samples of students and schools that are assumed to be nationally representative. The SLAD provides less depth, but includes a near census of all U.S. public schools, allowing for detailed comparisons across states and reducing reliance on sampling assumptions.

It is important to emphasize that every state uses a different standardized test, which makes it difficult in the SLAD to make direct comparisons between schools located in different states. The Education Trust reports use the state tests to calculate each school’s percentile ranking within the state. In some sense, this creates a common scale for all schools in the database, but it does not solve the problem. A school at the 40th percentile in North Dakota has a different level of achievement than one scoring at the same percentile in Montana. Therefore, when possible, the present study uses within-state analysis to make specific points.

Discussion and Analysis of Available Data

New Evidence about High Flyers

Table 1 reports the percentage of schools in each of four poverty/performance categories. The table uses the Education Trust definitions of high-poverty (50 percent of the school's students eligible for free and reduced lunches) and high-performance (the school is in the top-third of the state in either reading or math).

Only 16 percent of high-poverty schools are high-performing, compared with 54 percent of low-poverty schools. This means that low-poverty schools are three times more likely to be high performing than high-poverty schools. Notice also that 34 percent of all schools are high-poverty. Roughly 11.8 million students attend these schools. This large number reinforces the importance of reliably identifying high-flyers and learning from their effective practices.

Table 1 also provides the same information for schools with high levels of poverty *and* large portions of minority students. The results are even more disparate. In this case, only 10 percent of high-poverty-high-minority schools are high-performing, compared with 57 percent of low-poverty-low-minority schools, making them six times as likely to reach this high achievement level.³¹

Table 1: Poverty, Race, and Achievement in U.S. Public Schools

Category (ET definitions)	Number of Schools (% total sample)	% in each performance category	
		<i>Low</i>	<i>High</i>
Low-Poverty	40,830 (66%)	46	54
High-Poverty	21,234 (34%)	84	16
Total	62,064 (100%)		
Low-Poverty-Low-Minority	38,104 (61%)	43	57
High-Poverty-High-Minority	12,869 (21%)	90	10
Total	50,973 (82%)		

Accounting for regression to the mean

The above section uses the Education Trust definition of performance, which requires high-performance in either reading or math in the grade and year selected by Education Trust for analysis. This section considers the implications of this by providing analysis of the 18,365 schools in the “multiple scores” sample. Table 2 below shows the percentages of low- and high-poverty schools that are high-performing when various combinations of high scores are required. For instance, the first definition (1-1-1) refers to those schools that are high-performing in either year, either subject, and either grade. Because there are two subjects and two grades (four chances) and two years (giving four additional chances), each school has eight chances to get in the top-third just one time to become a high-performer. The 2-1-1 definition is somewhat more demanding, requiring that schools are high-performing in both years, in either subject, and in either grade. This definition requires more consistency than the one above. The degree of stringency

continues to increase up to the 2-2-2 definition, which requires schools to be high-performing in both years, both grades, and both subjects. Here, schools must have high test scores with a high degree of consistency.

Table 2 reports the percentage of low- and high-poverty schools that would be judged high-performing based on each of these definitions. Row 9 shows how schools would place on the basis of the Education Trust definition, which simply requires a school to be in the state's top-third in reading or math in one grade and year; it stands apart because Education Trust does not consider multiple years or multiple grades. The table's last row indicates the "erosion" of performance between the Education Trust definition of performance and the 2-2-2 definition. That is, it indicates what portion of schools drop off the high-performance list as criteria increasingly require consistent high-performance.

Table 2: Consistency of High-Performance

Row	Criteria			% high-poverty schools that are high-performing	% low-poverty schools that are high-performing	% high-poverty, high-minority schools that are high-performing	% low-poverty, low-minority schools that are high-performing
	<i>Years</i>	<i>Subjects</i>	<i>Grades</i>				
1	<i>1</i>	<i>1</i>	<i>1</i>	30.5	80.0	22.0	84.0
2	<i>2</i>	<i>1</i>	<i>1</i>	12.9	59.1	7.5	63.5
3	<i>1</i>	<i>2</i>	<i>1</i>	14.7	62.3	9.1	66.8
4	<i>1</i>	<i>1</i>	<i>2</i>	11.0	56.5	6.4	60.9
5	<i>2</i>	<i>2</i>	<i>1</i>	4.5	41.0	2.0	44.8
6	<i>2</i>	<i>1</i>	<i>2</i>	3.6	37.9	1.4	41.4
7	<i>1</i>	<i>2</i>	<i>2</i>	2.4	33.2	0.9	36.4
8	<i>2</i>	<i>2</i>	<i>2</i>	1.1	24.2	0.3	26.7
9	<i>Education Trust Definition</i>			<i>15.6</i>	<i>54.2</i>	<i>10.4</i>	<i>56.7</i>
10	<i>Erosion (from row 9 to row 8)</i>			<i>93 %</i>	<i>55 %</i>	<i>97 %</i>	<i>53 %</i>

The results suggest considerable decline in the percentages of schools that are high-performing as more consistency is required. As the earlier discussion of statistical noise indicated, any single measurement can be affected by external factors, and it is unlikely that a low-performing school that benefited from positive noise in one particular year would be able to sustain its high-performance ranking for two years. These results offer evidence supporting the earlier hypothesis that the designation of high-poverty, high-performing schools will be disproportionately affected by regression to the mean.

Indeed, the percentage of high-poverty schools achieving high performance declines from 15.6 percent using the Education Trust definition (Table 2, row 9) to 1.1 percent using the 2-2-2 definition (Table 2, row 8). This means that 93 percent of schools identified as high-flyers using the Education Trust approach are not high-flyers when consistency is required. The percentage also erodes for low-poverty schools, but not as much. The percentage of low-poverty schools achieving high performance declines from 54.2 percent using the Education Trust definition (Table 2, row 9) to 24.2 percent using the 2-2-2 definition (Table 2, row 8). This yields an erosion rate of 55 percent for low-poverty schools, considerably lower than the 93 percent found for high-poverty schools. The higher erosion rate for high-poverty schools confirms that the effect of regression to the mean plays a greater role in these schools, as explained earlier.

A further implication is that the probability of a high-poverty school reaching high-performance is much lower than the 2001 Education Trust report suggested. Recall that the results from Table 1, using the Education Trust definitions, suggested that low-poverty schools were three times as likely to be high-performing compared with high-poverty schools. Table 2 suggests that this number rises quickly when the performance

definition requires more consistency. The definition requiring the most consistency (row 8), which nearly eliminates the effect of regression to the mean, suggests that low-poverty schools are *22 times as likely to be high-performing*.³² The intuition behind this change is straightforward: both types of schools are less likely to be high-performing with the more restrictive definition, but the rate of erosion is higher for high-poverty schools. Therefore, the probability of high-poverty school reaching high-performance drops quickly compared with low-poverty schools.

Table 2 also provides evidence regarding schools that are high-poverty *and* high-minority. The initial portion of high-poverty, high-minority schools that are high performing is smaller than for high-poverty-only schools, consistent with the results in Table 1. The rate of erosion is also higher here, reaching 97 percent from the previous 93 percent. Further, the likelihood that a low-poverty-low-minority school is high performing is 89 times greater than for a high-poverty-high-minority school.

Accounting for proficiency definitions

This section tests whether the performance of high-poverty schools is sensitive to the cut score. Table 3 compares the math achievement for schools in Michigan and Florida, two states that reported results for each school using multiple cut scores.³³ First, the most noticeable differences are the levels of the cut scores across the two states. For these years, Michigan had relatively low cut scores, allowing high percentages of schools in all poverty categories to reach high performance, even with the highest cut score. For example, 85 percent of Michigan's high-poverty elementary schools reached the lowest cut score. This suggests that Michigan's cut scores are closer to "cut score 1" in Figure 1.

Florida, in contrast, uses more cut scores and has a much wider range of students passing. The lowest cut score appears more like “cut score 1” in Figure 1, while the highest cut score is more like “cut score 2.” These definitions are of course somewhat arbitrary in both states and the point here is simply to illustrate the influence of these choices.

The last column most clearly illustrates the point that the achievement gap appears largest when using cut scores that are closest to the intersection of the test score distributions for advantaged and disadvantaged students. In Michigan, the difference in the percentages of students passing between low- and high-poverty schools is relatively low with the lowest cut score, but the gap widens moving to the highest cut score—that is, closer to the intersection of the two distributions of scores shown in Figure 1.

A similar pattern is observed in Florida, when shifting from the lowest to the middle cut score. Interestingly, the difference between low- and high-poverty schools decreases again when shifting from the middle to the highest cut score. The apparent reason is that the middle cut score is near the intersection of the advantaged and disadvantaged student distributions—where the gap is greatest. Shifting from the middle to the highest standard therefore shifts the cut score away from the intersection of the test score distributions for advantaged and disadvantaged students. Thus, the results from Florida also reinforce Rothstein’s point, although the point is made somewhat differently because of the wide range of scores used in that state.

Table 3: Role of the State Proficiency Definitions, Individual States

Cut Scores	Average % of students reaching cut score in high-poverty schools (a)	Average % of students reaching cut score in low-poverty schools (b)	Difference (b)-(a)
Michigan			
Elementary			
Lowest Cut Score	85.0	94.4	9.4
Highest Cut Score	63.2	80.8	17.6
Middle			
Lowest Cut Score	70.2	89.2	19.0
Highest Cut Score	38.7	67.3	28.6
Florida			
Elementary			
Lowest Cut Score	65.6	83.4	17.8
Middle Cut Score	34.7	56.3	21.6
Highest Cut Score	14.3	28.8	14.5
Middle			
Lowest Cut Score	56.5	79.2	22.7
Middle Cut Score	34.6	59.8	25.2
Highest Cut Score	12.6	28.6	16.0

NCLB Revisited

Consider again the connection between these results and NCLB. First, the new federal law is based on the lowest state cut scores, such as those in Table 3 for Florida

and Michigan. These definitions have been criticized because they are somewhat arbitrary and because the law indirectly encourages states to create definitions of proficiency that are easy for students to reach. These concerns are all valid. And the present discussion adds another to the list: NCLB makes the achievement gap look smaller than it is in reality. For instance, the results in Table 3 for Michigan suggest that the gap between low- and high-poverty schools is as low as 9.4 percent points. While not directly comparable to the achievement gaps described in the above “Background” section, the number does give an impression that the problem is relatively small. In short, NCLB may have the effect of reducing the gap simply by defining it away.

But even this problem should not distract us from the more fundamental flaw—NCLB assumes that schools are solely responsible for student achievement. This assumption is misguided, given the strong evidence regarding the role of poverty and students’ home and community environments.

Recommendations

This study has re-analyzed data from recent reports that purport to show large numbers of “high-flying” schools, which they then use as evidence to suggest that overcoming social and economic disadvantage is relatively easy. Three flaws have been identified.

First, Table 2 shows that, after accounting for regression to the mean and requiring consistent high performance, the number of high flyers seems quite small. For example, a low-poverty-low minority school is 89 times more likely to be high-performing than one that is high-poverty-high-minority.

Second, the logic of the high flyer argument is also flawed. Even if there were large numbers of high flyers, this would say relatively little about the role of social and economic disadvantages. Much better evidence on this question comes from research on kindergartners whose achievement cannot reasonably be attributed to schools. That evidence shows that students start school almost as far behind as when they finish school. While schools must take some responsibility for these gaps, the evidence shows clearly students' disadvantages are the primary cause behind the achievement gaps.

Third, the notion that schools are solely responsible for educational inequity is a basic, but entirely misguided, assumption of state and federal education policy, including *No Child Left Behind*. The related emphasis on school accountability, most commonly based on cut scores, and the federal focus on minimum proficiency together work to distort the size of a critical achievement gap that needs serious attention and intervention. Instead, as shown in Table 3, it is more likely that NCLB will “reduce” the achievement gap simply by redefining it.

It is therefore recommended that:

1. Policy makers continue the recent focus on measurable student outcomes, such as test scores, but redesign policies to hold educators accountable only for those factors within their control;
2. Policy makers take a comprehensive approach to school improvement that starts in schools, but extends into homes and communities, addressing basic disadvantages caused by poverty; and

3. All educational stakeholders acknowledge that educational inequity is caused by problems in both schools *and* communities—and avoid trying to blame the problem on schools alone.

Acknowledgements

This research was produced with editorial assistance from the Arizona State University Education Policy Studies Laboratory. The author wishes to thank Debbi Harris, Patricia Hinchey, Alex Molnar, Kevin Welner, and anonymous referees for their valuable comments. Richard Kahlenberg, Patrick McEwan, Lawrence Mishel, and Richard Rothstein also provided valuable comments in earlier versions. Donald McLaughlin at the American Institutes for Research provided useful advice regarding some of the data. Jian Gao provided valuable research assistance. All remaining errors are the author's responsibility.

Notes & References

¹ The percentiles are converted from standard deviation form using the standard normal distribution. For long-term evidence on the gaps using the SAT, ACT, MCAT, LSAT, and GRE, see:

Camara, W. & Schmidt, E. (1999). *Group differences in standardized testing and social stratification* [College Board Rep. No. 99-5]. New York, NY: College Entrance Examination Board.

For evidence regarding the NAEP, see:

Phillips, C. & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.). *The Black-White Test Score Gap* (pp. 229-272). Washington, DC: Brookings.

² Regarding the achievement gap in recent decades, see:

Harris, D.N. & Herrington, C.D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2), 209-238.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, 31(1): 3-12.

Regarding earlier decades, see:

Phillips, C. & Ralph, J. (1998). Does the black -white test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.). *The Black-White Test Score Gap* (pp. 229-272). Washington, DC: Brookings.

³ Regarding the nature of these disadvantages, see:

Rothstein, R. (2004). *Class and Schools*. Teacher's College Press: New York.

⁴ See:

Brooks-Gunn & Duncan (1997). The effects of poverty on children. *The Future of Children*, 7(2), 55-71.

⁵ See:

Duncan G. J. & Brooks-Gunn, J. (2000). Family poverty, welfare reform, and child development. *Child Development*, 71(1), 188-196.

⁶ See:

Conger, K.J., Conger, R., & Scaramella, L.V. (1997). Parents, siblings, psychological control and adolescent adjustment. *Journal of Adolescent Research*, 12(1), 113-138.

⁷ See:

Brooks-Gunn, J., Klebanov, P.K., Smith, J., Duncan, G.J., & Lee, K. (2003). The black-white test score gap in young children: Contributions of test and family characteristics. *Applied Developmental Science* 7(4) 239-252.

⁸ For evidence regarding kindergartners, see:

Lee, V.E. & Burkham, D.T. (2002). *Inequality at the Starting Gate*. Washington, DC: Economic Policy Institute.

⁹ For example, some argue that housing segregation is almost completely immune from government policy, see:

Thernstrom, A. & Thernstrom, E. (2002). *No Excuses: Closing the Racial Gap in Learning*. New York: Simon & Shuster.

¹⁰ The debate about the role of resources such as class size and teacher salaries is an old one. For example, some conducted extensive literature reviews on the topic. See:

Greenwald, R., Hedges, L.V., & Laine, R.D., (1996). The effect of school resources on student achievement. *Review of Educational Research*, 66(3), 361-396.

Hanushek, E.A., (1996). A more complete picture of school resource policies. *Review of Educational Research*, 66(3), 397-409.

A separate review finds mixed evidence regarding resource effects, but that the effects are apparently positive and larger for disadvantaged students. See:

Harris, D.N. & Herrington, C.D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2), 209-238.

¹¹ One recent study finds evidence that some types of standards and accountability may improve equity, but the authors also point out the inherent difficulties of identifying the effects of these policies because test scores become distorted when such policies are put in place; that is, changes in test scores after the policies are put in place may not reflect genuine changes in learning. See:

Harris, D.N. & Herrington, C.D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, 112(2), 209-238.

¹² This point has been made by others. See, for example:

Welner, K.G. (2005). Can irrational become unconstitutional? NCLB's 100% presuppositions. *Excellence and Equity in Education*, 38, 171-179.

¹³ In addition to ignoring student learning gains, the incentives in NCLB apply only to schools, not to families or community institutions that also affect student learning. Also, while NCLB involved an increase in federal funding, the federal share is very small and therefore is hardly noticeable to schools. By doing little to increase school resources, the law implies that current resources are not being used efficiently.

¹⁴ The citation for the 2001 Education Trust report is: Jerald, C. (2001). *Dispelling the myth: Revisited*. Washington, DC: The Education Trust.

¹⁵ A preliminary version of the analysis in the present study was presented by the author to representatives of the Education Trust prior to the March, 2002 follow-up report from Education Trust.

¹⁶ The quotation and the tables being referred to are in:

Education Trust (March, 2002). *Dispelling the myth revisited: Additional tables*. Washington, DC: Author.

Additional tables were published in:

Education Trust (May, 2002). *Dispelling the myth . . . over time*. Washington, DC.

¹⁷ Quoted in:

Nathan, J., (2002, January 6,) Successful schools share many traits. *St. Paul Pioneer Press*.

¹⁸ Here are several other examples of how the Education Trust report has been interpreted in the press: "The Education Trust's new report proves beyond dispute that poor and minority children can achieve at high levels – provided educators have high expectations for these students." See:

Stanton, B. (2001, December 16). Challenged schools score big victories: Poor, minority students buck trends and excel. *Denver Post*.

"It's further proof that all kids can learn . . . There are no excuses," Dudley Williams, Stamford, CT, Board of Education. See:

Christoffersen, J. (2001, December 29). Study identifies strong schools with poor students. *USA Today*.

"This report shows that we don't have to accept low achievement." See:

Nathan, J., (2002, January 6) Successful schools share many traits. *St. Paul Pioneer Press*.

¹⁹ The citation for the Heritage report is:

Carter, S. (2000). *No Excuses: Lessons from 21 High-Performing, High-Poverty Schools*. Washington, DC: The Heritage Foundation.

Another example of this approach involves the so-called 90-90-90 schools in which 90 percent of students receive free or reduced price lunches, 90 percent of their students belong to ethnic minority groups, and 90 percent of their students meet or exceed state academic standards. See <http://www.makingstandardswork.com/index.htm>.

²⁰ In addition to nine of the 21 schools having admission requirements that could exclude students who have received low test scores, a tenth school, Kew Elementary School in Inglewood, California, housed a gifted and talented program that also may have had admission requirements. Of the remaining schools, four required tuition, suggesting that their parents may have access to financial resources not reflected in free and reduced lunch eligibility. The fact that low-income families are willing to find such resources also suggests that these parents value education more than the typical parent, which may also mean that they have exceptional skills and experiences required to help their children excel. Another of the remaining schools is located in Cambridge, Massachusetts – home of two of the most prestigious universities in the nation, Harvard and MIT. Large portions of the students in these schools were children of highly educated graduate students whose low incomes were only temporary circumstances related to their education programs. These parents were clearly able to provide resources necessary for their children to succeed. Another of the remaining schools, 14th Avenue Elementary School in New Jersey, had a pupil-teacher ratio of just 14. Portland Elementary School in Arkansas has only 152 students, far fewer than the average school, and is located in a rural area. Also, the Newberry Elementary in Detroit, which had after-school programs four days per week, plus a “summer learning academy,” providing a substantial amount of additional time for learning. A similar discussion of the Heritage report is provided by Rothstein (2004).

²¹ For example, see:

Kane, T. & Staiger, D. (2001). *Improving school accountability measures* [NBER Working Paper #8156]. National Bureau of Economic Research: Cambridge, MA.

²² This also suggests that schools with more students will experience less noise and therefore be less affected by regression to the mean. Further, because elementary schools tend to have fewer students taking any given test, regression to the mean especially influences the average scores of elementary schools relative to middle and high schools.

²³ This quotation is on page 15 in:

Rothstein, R. (2004). *Class and Schools*. Teacher’s College Press: New York.

²⁴ This quotation is on page 89 in:

Rothstein, R. (2004). *Class and Schools*. Teacher’s College Press: New York.

²⁵ One assumption is that only the mean of the distribution differs between advantaged and disadvantaged students. (Notice in the figure that the two distributions have identical “bell” shapes.) It is also important to point out that these shifts in the relative effects on the two groups are larger when the distributions are more distinct from one another. If the distributions largely overlapped, this would not be a significant issue.

²⁶ Free and reduced lunch (FRL) eligibility is the only measure of student economic well being available in this analysis. Eligibility for reduced priced lunches requires income less than 185 percent of the federal poverty rate. Free lunch eligibility requires income less than 130 percent of the poverty rate. Four states have reduced lunch data in the SLAD, but not free lunch data (Arizona, Illinois, Tennessee, and Washington). Therefore, reduced lunch eligibility was imputed based on the relative number of free and reduced lunch students in the average state. This was done by

regressing reduced lunch data on free lunch and the other variables for the states that had both, using the variables in specification (2) of table 1d (except poverty). I then used this equation to estimate the number of reduced lunch students in the states that had only free lunch. Students eligible for reduced lunches are added to those eligible for free lunches, even though each type reflects a different level of family income.

- ²⁷ The SLAD includes more data than is used by the Education Trust. The ET reports are unclear regarding what specific data they used. Nevertheless, I was able to closely replicate their analysis.
- ²⁸ Iowa had no statewide testing program during the specified years. West Virginia and South Dakota reported only one score in each grade, which was insufficient to calculate performance by any of the definitions considered in this paper, including the ET definition.
- ²⁹ The present study uses raw scores and scale scores when possible because these are less sensitive to small random noise. More information on the test scores used in each state is available from the author upon request. The Education Trust report does not include discussion of how the scores were chosen for each state.
- ³⁰ School level determinations are made as follows. The first step for each school is to identify the highest grade in which a test score is reported. If this highest grade was six or lower, the school is categorized as “elementary”; when the highest grade is eight or lower, the school is labeled “middle”; when the highest grade above eight, the school is considered “high.” The next step is to identify the minimum grade at which a score is reported. If a middle school has a grade less than six, then a separate variable was created identifying the school as “multi-level” and given the value of one. The same is true for high schools reporting scores in grades less than nine.
- ³¹ Note that this table includes only 82 percent of all schools, excluding those that are high in poverty, but low in minorities, and schools that are low in poverty, but high in minorities.
- ³² The probability of a high-poverty school being high-performing is 24.2. The probability of a low-performing school being high-performing is 1.1. Dividing these two yields 22.
- ³³ For Michigan, I used mathematics achievement in grade four and seven. For Florida, I used mathematics achievement in grades five and eight. High school results are excluded for sake of space.